2011-11-21

Design of In uenza Surveillance Networks

Dimitrov, Nedialko B.

Monterey, California: Naval Postgraduate School.

Design of Influenza Surveillance Networks

Nedialko B. Dimitrov[1] Samuel Scarpino[2]
Lauren Ancel Meyers[2]

Naval Postgraduate School[1]

The University of Texas at Austin[2]

November 21, 2011

Background: Influenza

1. New strains recurringly create global pandemics
2. Most recent, 2009 swine-origin influenza

Background: Influenza

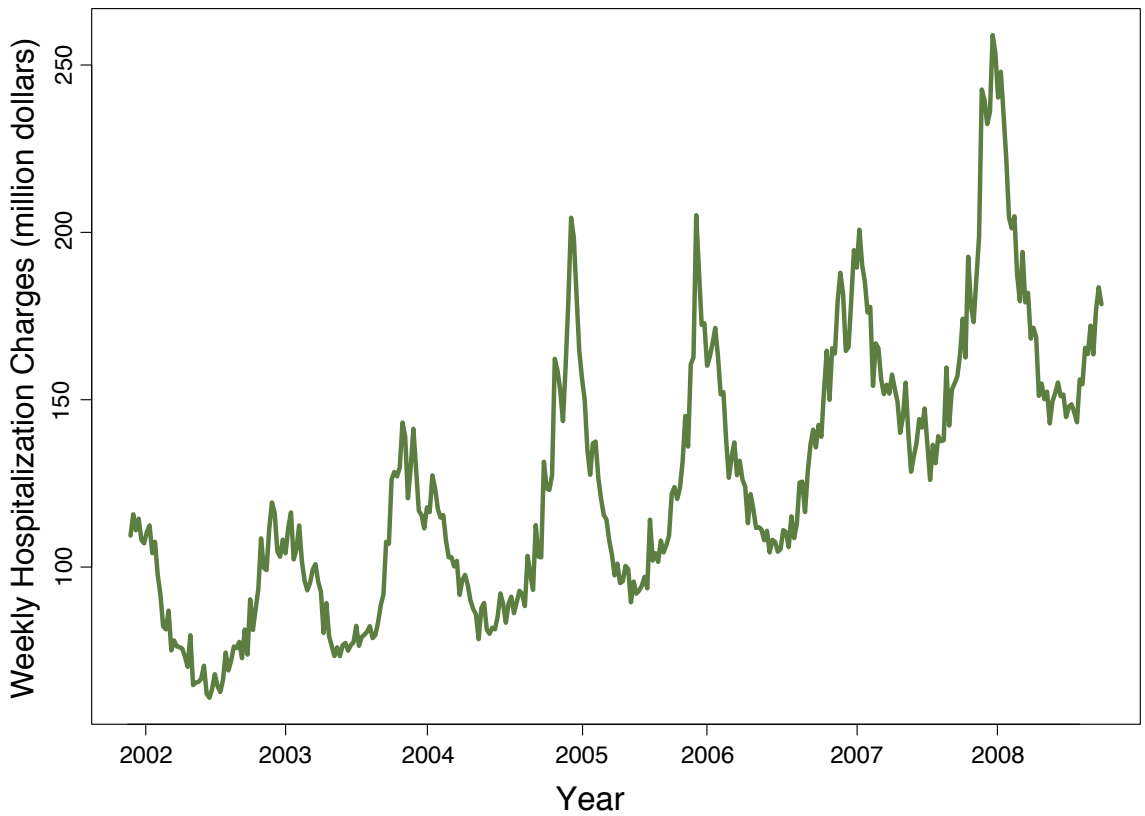
Leading causes of death in the U.S. [CDC]

1. Heart disease
2. Cancer
3. Stroke
4. Chronic lower respiratory diseases
5. Accidents
6. Alzheimer's disease
7. Diabetes
8. Influenza and Pneumonia
9. Nephritis, nephrotic syndrome, and nephrosis
10. Septicemia

Background: Influenza

In Texas:

- 23-37% of respiratory related hospitalizations
- \$9.5 billion in hospitalization bills in 2008, $\approx$ 1% GDP

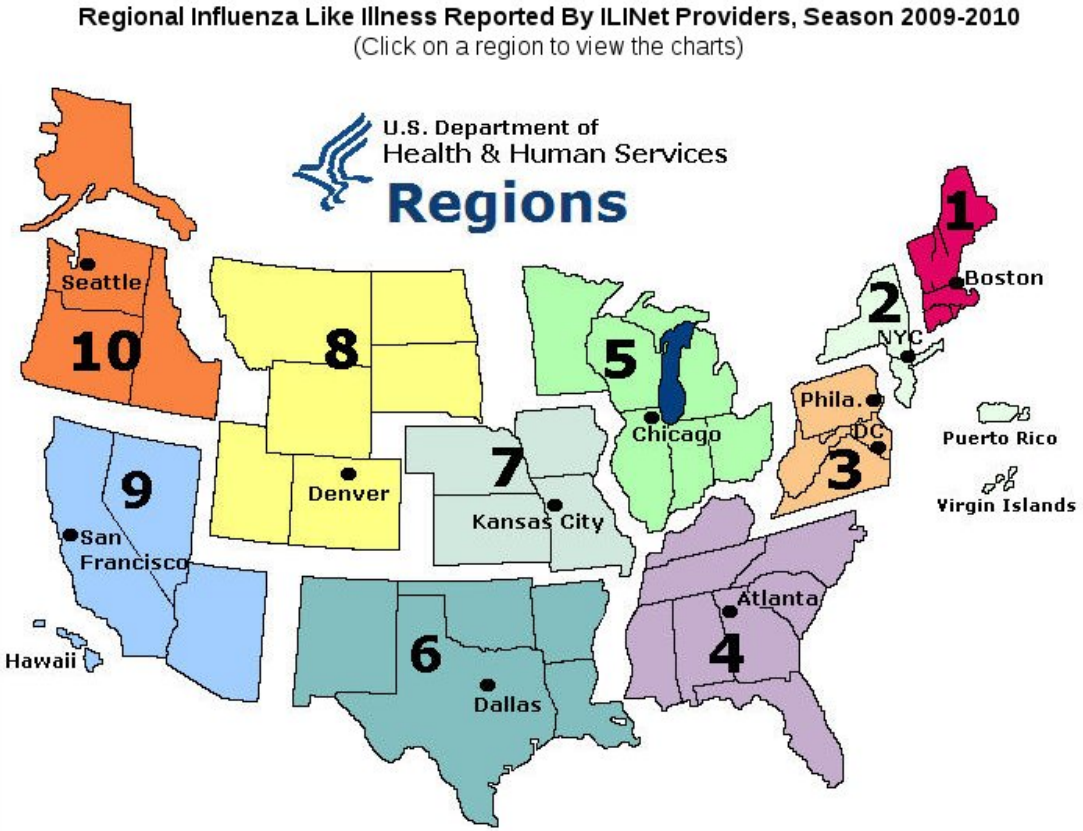


Background

- Hospitals would like warning before an inflow of patients
- CDC's Influenza-like-illness Network (ILINet):
  - A few primary healthcare providers report weekly:
    - Total number of patients seen
    - Number of patients with influenza-like-illness for five age groups (0-4, 5-24, etc.)

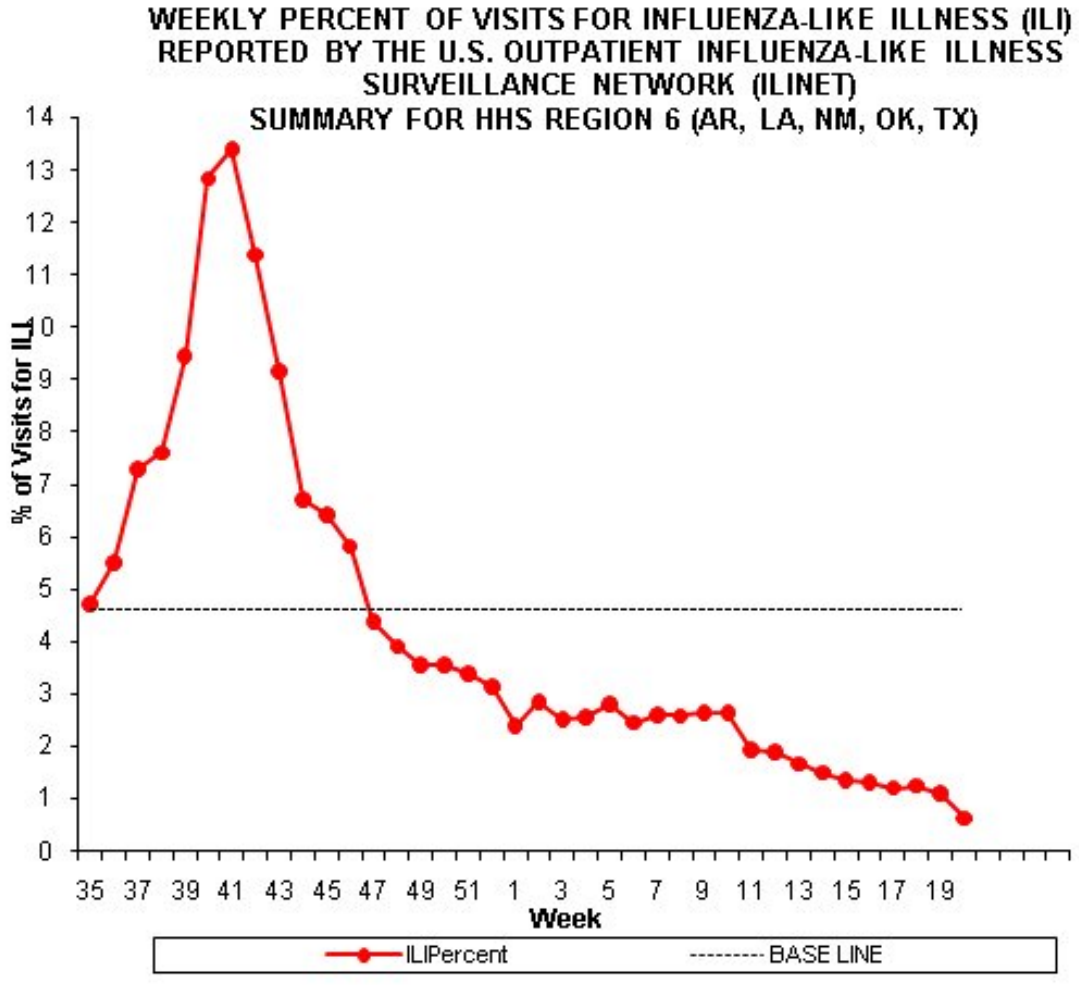
Background

For example, CDC's publicly available aggregates of ILINet:



Background

For example, CDC's publicly available aggregates of ILINet:



Problem

- Each state is responsible for recruiting ILINet providers
- Texas legislature asked Department of State Health Services:
  - Does ILINet work?
  - How do we design a better ILINet?

Problem

Subproblems:

1. Thousands of possible providers?
2. Noise in the provider reports?
3. What is the objective function?
4. What algorithm to select providers?
5. How do we compare to existing networks and methods?

Data

ILINet reports:

- Zip code for each provider.
- Provider's reports over time.

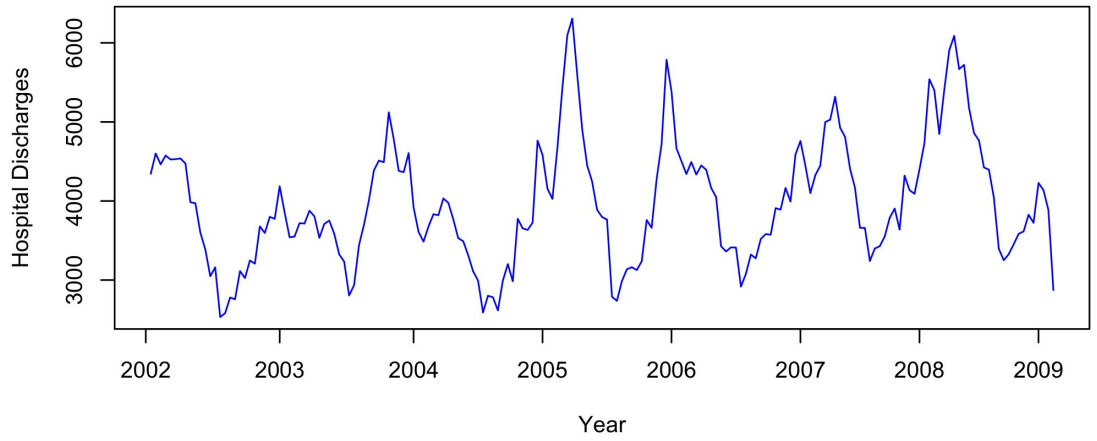
Hospital discharge data:

- Home zip code of each patient discharged in TX.
- Patient's diagnosis codes.

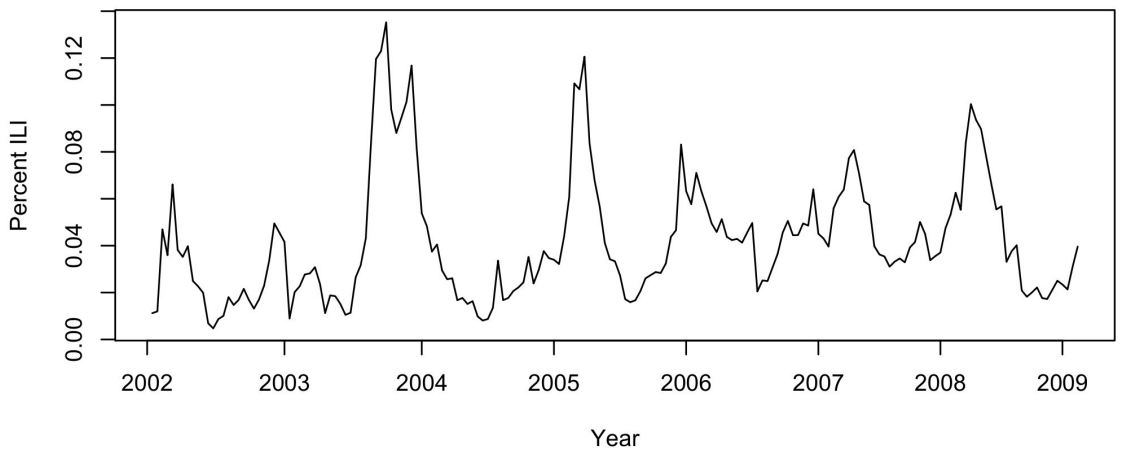
Overlap Aug 2001 to May 2008.

Data: Example

From hospital discharge records:



From ILINet reports:



A Pool of Mock Providers

Subproblems:

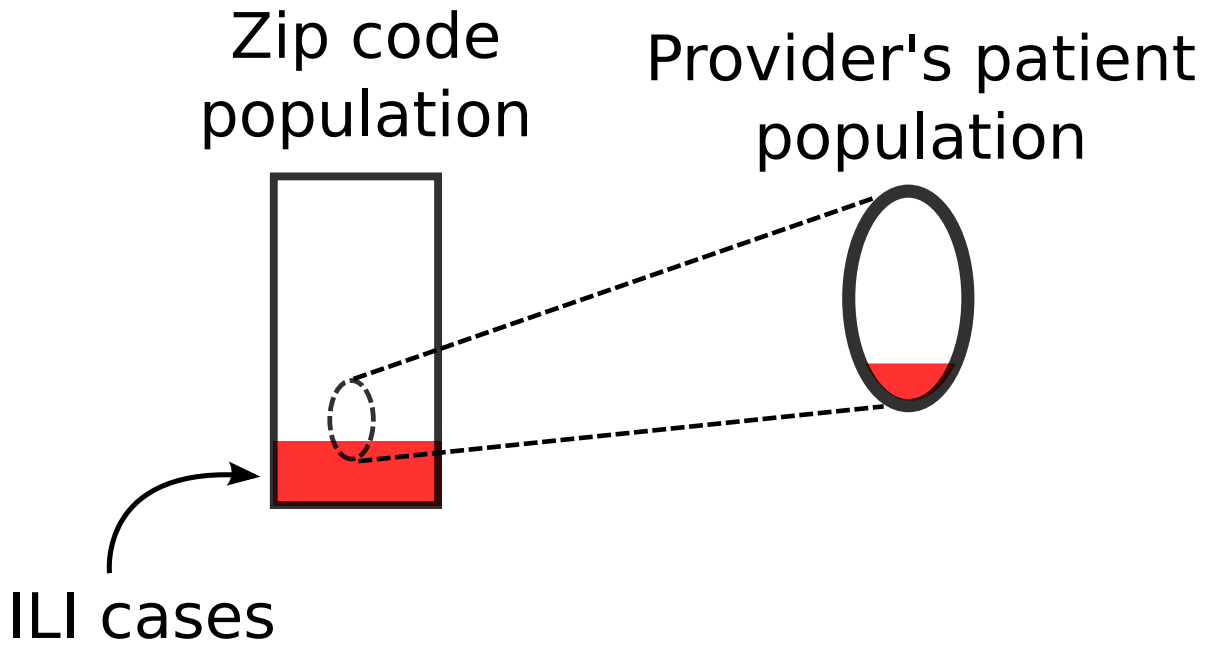
1. Thousands of possible providers?
2. Noise in the provider reports?

Approach:

- Derive provider noise profiles from existing providers
- Generate mock providers using the noise profiles

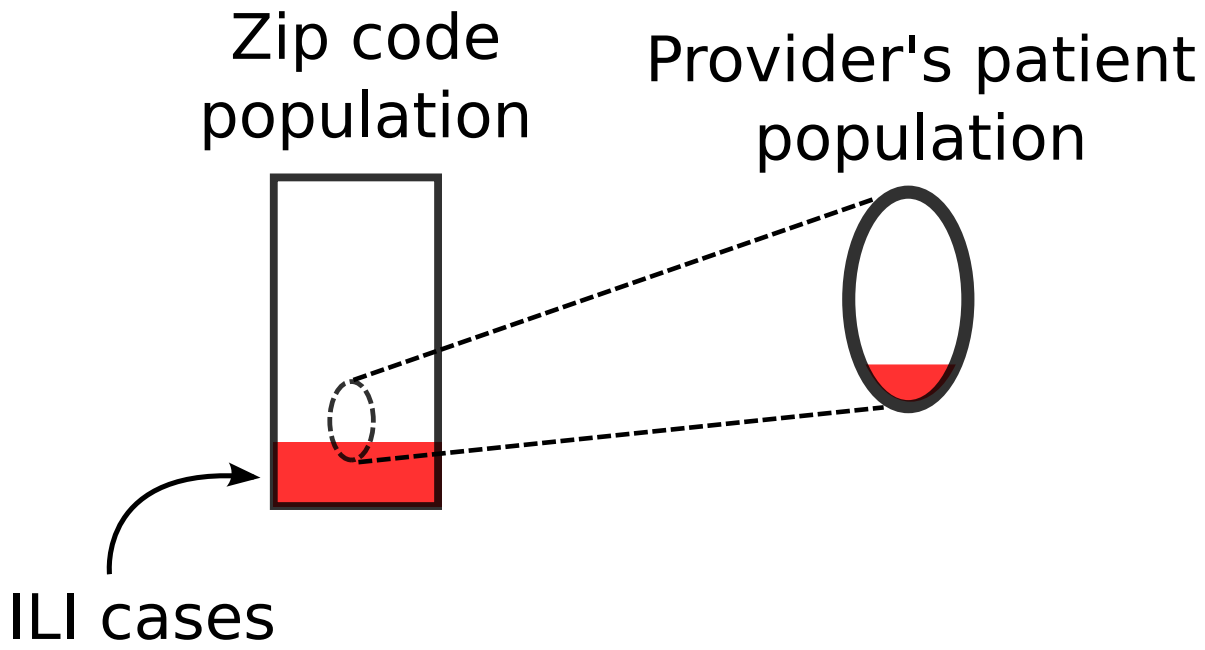
A Pool of Mock Providers

Provider model concept:



A Pool of Mock Providers

Provider model concept:



1. Data does not have fraction of ILI in each zip code
2. Noise in the reports

A Pool of Mock Providers

To estimate fraction of ILI in a zip code:

- Know number of hospitalizations in the zip code
- Know the hospitalization rate (previous studies)
- Gives Pr[hospitalizations | ILI cases]
- Bayes formula gives Pr[ILI cases | hospitalizations]

A Pool of Mock Providers

To estimate fraction of ILI in a zip code:

- Know number of hospitalizations in the zip code
- Know the hospitalization rate (previous studies)
- Gives Pr[hospitalizations | ILI cases]
- Bayes formula gives Pr[ILI cases | hospitalizations]

We use the expectation as the number of ILI cases in the zip

A Pool of Mock Providers

Two salient noise characteristics:

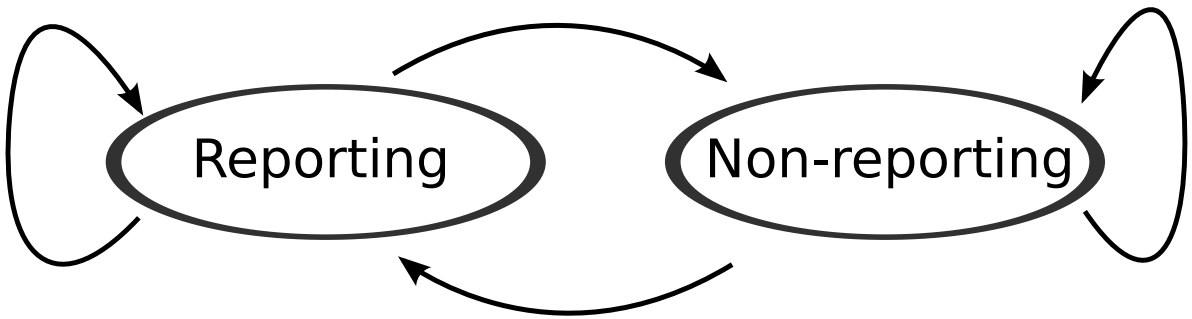
1. Report availability
2. Error in report data

A Pool of Mock Providers

Two salient noise characteristics:

1. Report availability
2. Error in report data

Report availability:

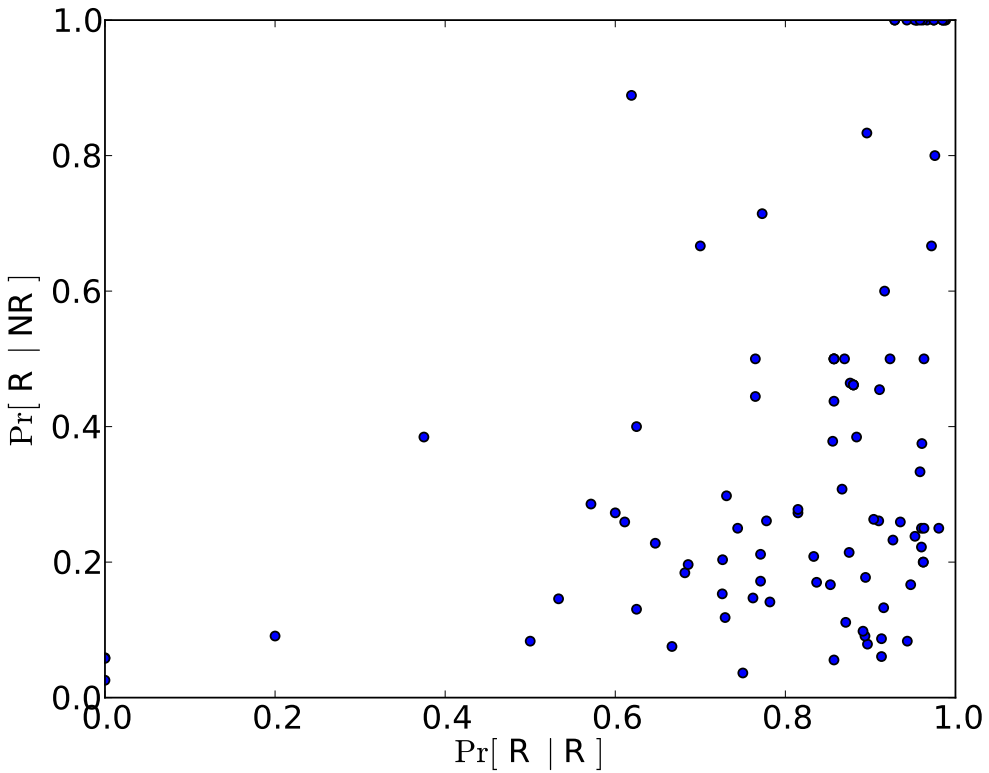


A Pool of Mock Providers

Two salient noise characteristics:

1. Report availability
2. Error in report data

Report availability:



A Pool of Mock Providers

Two salient noise characteristics:

1. Report availability
2. Error in report data

Error in report data:

$$\text{Provider-report}(i) = c_0 + c_1 \text{Percent-ILI}(i) + N(0, \sigma^2),$$

A standard regression noise model

A Pool of Mock Providers

Two salient noise characteristics:

1. Report availability
2. Error in report data

Gives an existing provider noise profile:
(transition probabilities, regression constants)

A Pool of Mock Providers

Two salient noise characteristics:

1. Report availability
2. Error in report data

Gives an existing provider noise profile:
(transition probabilities, regression constants)

To generate a mock provider for a zip code:

1. Select a uniformly random noise profile
2. Generate reports based on the profile

Gives a pool ≈ 2000 mock providers, one for each TX zip

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Approach:

- Use hospital discharge records for objective
- Exploit submodular objective [Das, Kempe; 08]

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function (first try):

$\mathbf{G}$ – goal time series, state-wide hospitalizations

P – mock provider pool

S – subset of providers selected

$$R^2(\mathbf{G}, S) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i})}{\mathrm{Var}(\mathbf{G})}$$

$$\max_{S \subseteq P} R^2(\mathbf{G}, S)$$

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function (first try):

$\mathbf{G}$ – goal time series, state-wide hospitalizations

P – mock provider pool

S – subset of providers selected

$$R^2(\mathbf{G}, S) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i})}{\mathrm{Var}(\mathbf{G})}$$

$$\max_{S \subseteq P} R^2(\mathbf{G}, S)$$

Submodular: [Das, Kempe; 08]

Submodular Functions

Def: Submodular

$$f : 2^P \rightarrow \Re$$
$$f(A + x) - f(A) \geq f(B + x) - f(B)$$

for $A \subseteq B$ and $x \notin A, B$. (diminishing returns)

Submodular Functions

Def: Submodular

$$f : 2^P \rightarrow \Re$$
$$f(A + x) - f(A) \geq f(B + x) - f(B)$$

for $A \subseteq B$ and $x \notin A, B$. (diminishing returns)

Key property:

$$f(A) \leq f(B) + \sum_{x \in A - B} [f(B + x) - f(B)]$$

for all sets A and B.

Submodular Functions

Def: Submodular

$$f : 2^P \rightarrow \Re$$
$$f(A + x) - f(A) \geq f(B + x) - f(B)$$

for $A \subseteq B$ and $x \notin A, B$. (diminishing returns)

Key property:

$$f(S^*) \leq f(B) + \sum_{x \in S^* - B} [f(B + x) - f(B)]$$

for all sets S^* and B. (gives optimality gaps)

Submodular Functions

Greedy Algorithm:

- Start with S empty.
- Loop:
  - Find x that maximizes $f(S + x) - f(S)$
  - Add x to S

Submodular Functions

Greedy Algorithm:

- Start with S empty.
- Loop:
  - Find x that maximizes $f(S + x) - f(S)$
  - Add x to S

Produces approximately optimal solutions: factor $1 - \frac{1}{e}$.

- Short proof
- Uses key optimality gap property

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function (first try):

$\mathbf{G}$ – goal time series, state-wide hospitalizations

P – mock provider pool

S – subset of providers selected

$$R^2(\mathbf{G}, S) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i})}{\mathrm{Var}(\mathbf{G})}$$

$$\max_{S \subseteq P} R^2(\mathbf{G}, S)$$

Submodular: [Das, Kempe; 08]

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function:

G – goal time series, state-wide hospitalizations

P – mock provider pool

S – subset of providers selected

$\tilde{\xi}$ – noise profile of provider pool

$$R^2(\mathbf{G}, S, \tilde{\xi}) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i}(\tilde{\xi}))}{\mathrm{Var}(\mathbf{G})}.$$

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})]$$

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function:

G – goal time series, state-wide hospitalizations

P – mock provider pool

S – subset of providers selected

$\tilde{\xi}$ – noise profile of provider pool

$$R^2(\mathbf{G}, S, \tilde{\xi}) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i}(\tilde{\xi}))}{\mathrm{Var}(\mathbf{G})}.$$

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})]$$

Convex combination of submodular functions also submodular

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function:

G – goal time series, state-wide hospitalizations
P – mock provider pool
S – subset of providers selected
$\tilde{\xi}$ – noise profile of provider pool

$$R^2(\mathbf{G}, S, \tilde{\xi}) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i}(\tilde{\xi}))}{\mathrm{Var}(\mathbf{G})}.$$

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})]$$

Three operators: $\max_S$, $E_{\tilde{\xi}}$, $\min_\alpha$

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

Objective function:

$\mathbf{G}$ – goal time series, state-wide hospitalizations
P – mock provider pool
S – subset of providers selected
$\tilde{\xi}$ – noise profile of provider pool

$$R^2(\mathbf{G}, S, \tilde{\xi}) = \frac{\mathrm{Var}(\mathbf{G}) - \mathrm{Var}(\mathbf{G} - \sum_{i \in S} \alpha_i \mathbf{P_i}(\tilde{\xi}))}{\mathrm{Var}(\mathbf{G})}.$$

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})]$$

Three operators: $\max_S$, $E_{\tilde{\xi}}$, $\min_\alpha$ . . . computationally difficult.

Optimization

Subproblems:

3. What is the objective function?
4. What algorithm to select providers?

R^2 and submodularity give:

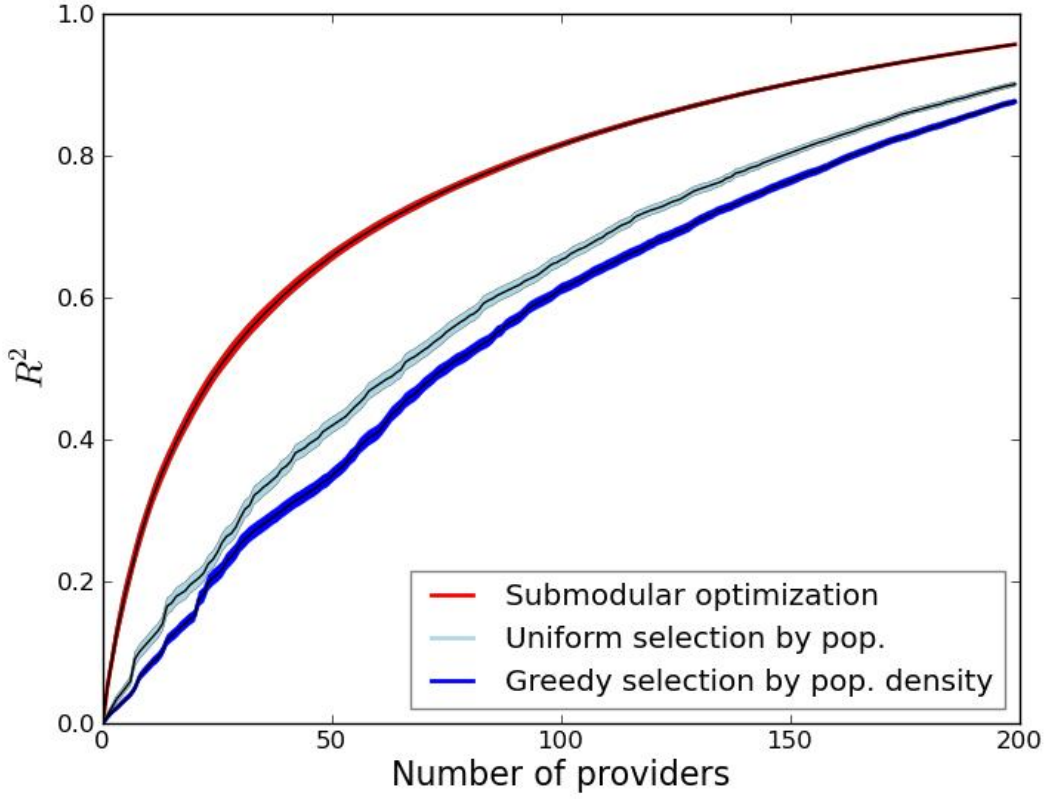
- A greedy algorithm for selection
- An optimality gap guarantee

Results

Subproblems:

5. How do we compare to existing networks and methods?

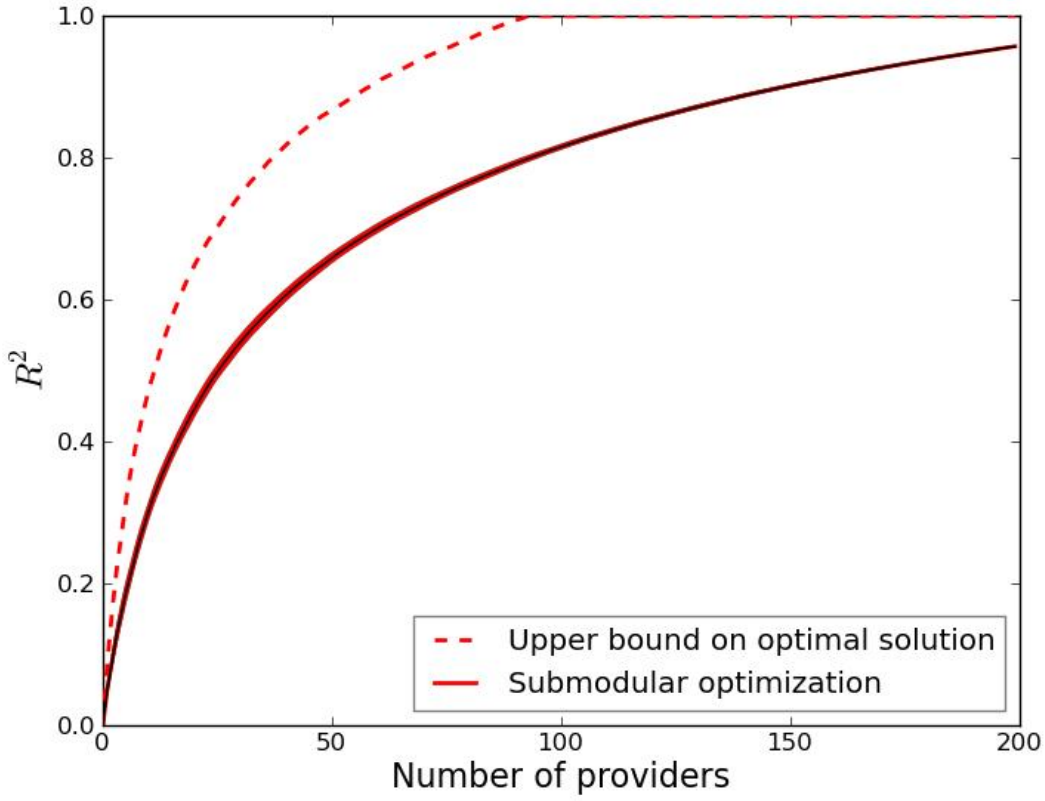
Results



Comparison with "naive" strategies:

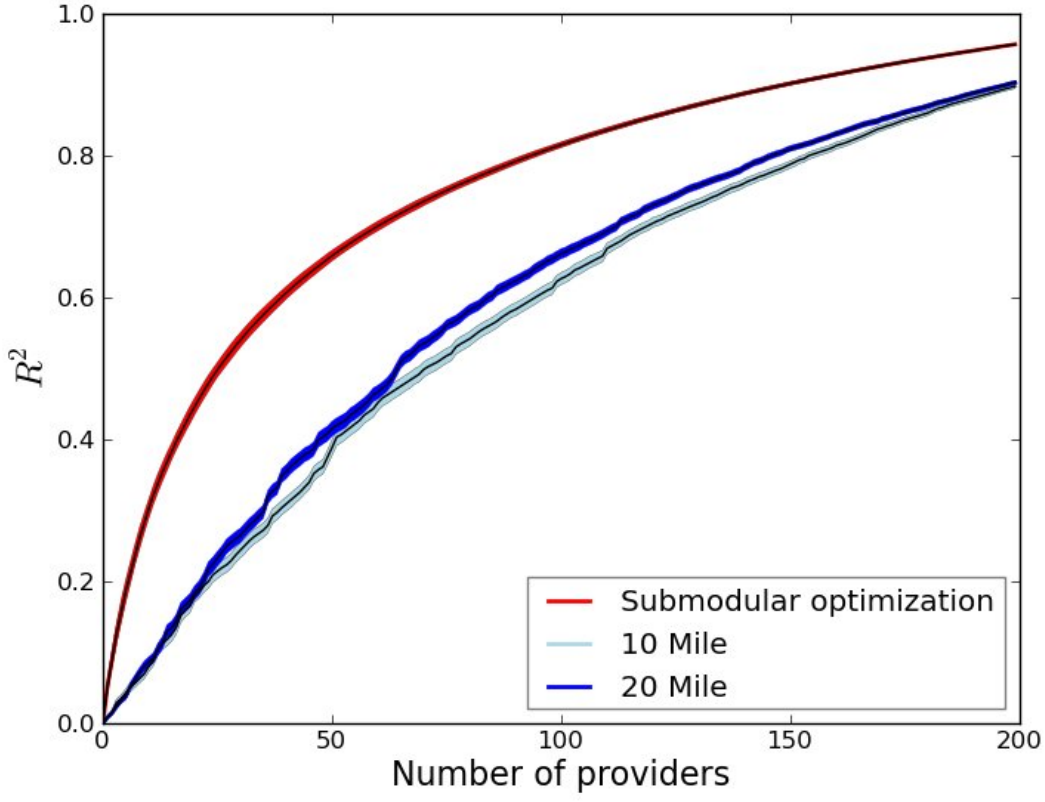
- Uniformly random selection models "open call" for providers
- Greedy selection models focusing on population centers

Results



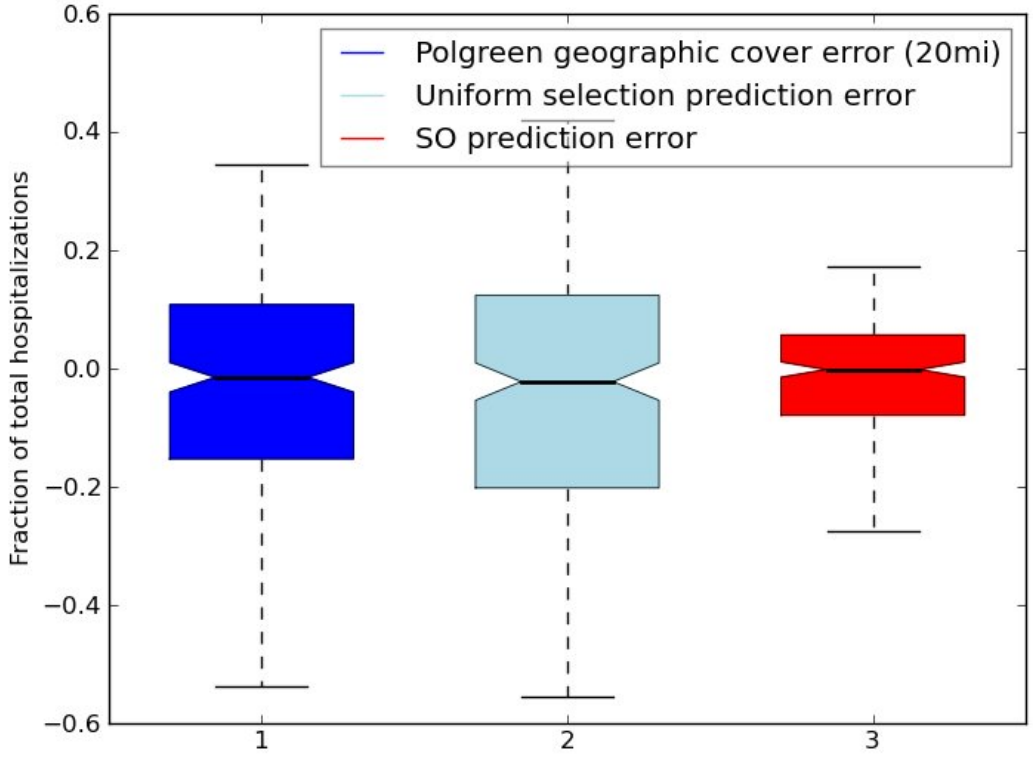
Comparison with optimal.

Results



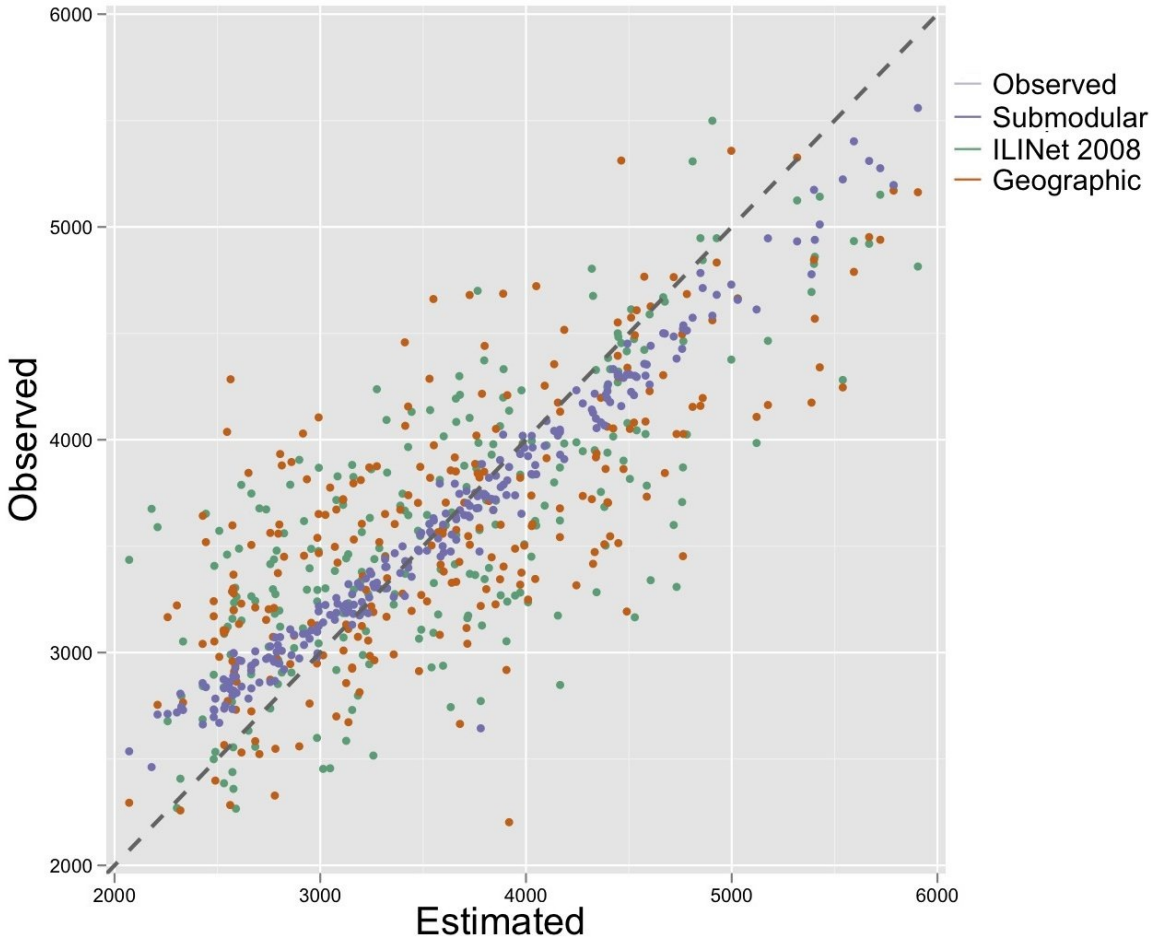
Comparison with [Polgreen et al. 09] (geographic covering).

Results



Prediction errors with 50 providers

Results



Scatter plot of in-sample predictions.

Results: Out-of-sample Validation

1. Create ILINet using data from 2001-2007
2. Extract α_i regression coefficients from 2001-2007
3. Create prediction from 2008 provider reports

Results: Out-of-sample Validation

1. Create ILINet using data from 2001-2007
2. Extract α_i regression coefficients from 2001-2007
3. Create prediction from 2008 provider reports

Models:

- Historical data to create network
- Historical data to create prediction function
- Real-time provider reports, to create prediction

Results: Out-of-sample Validation

1. Create ILINet using data from 2001-2007
2. Extract α_i regression coefficients from 2001-2007
3. Create prediction from 2008 provider reports

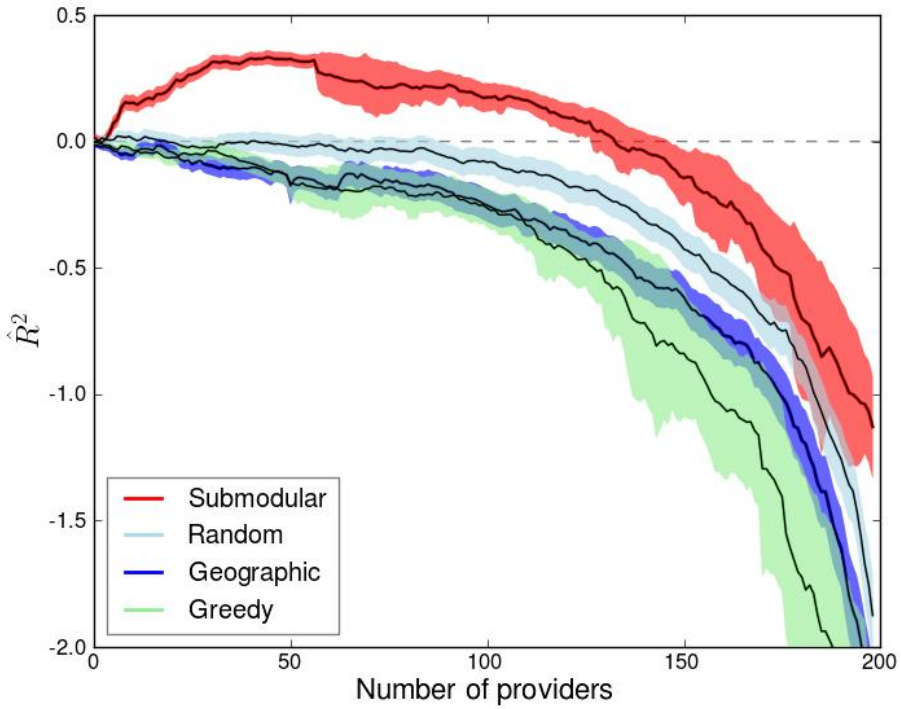
Models:

- Historical data to create network
- Historical data to create prediction function
- Real-time provider reports, to create prediction

Measure:

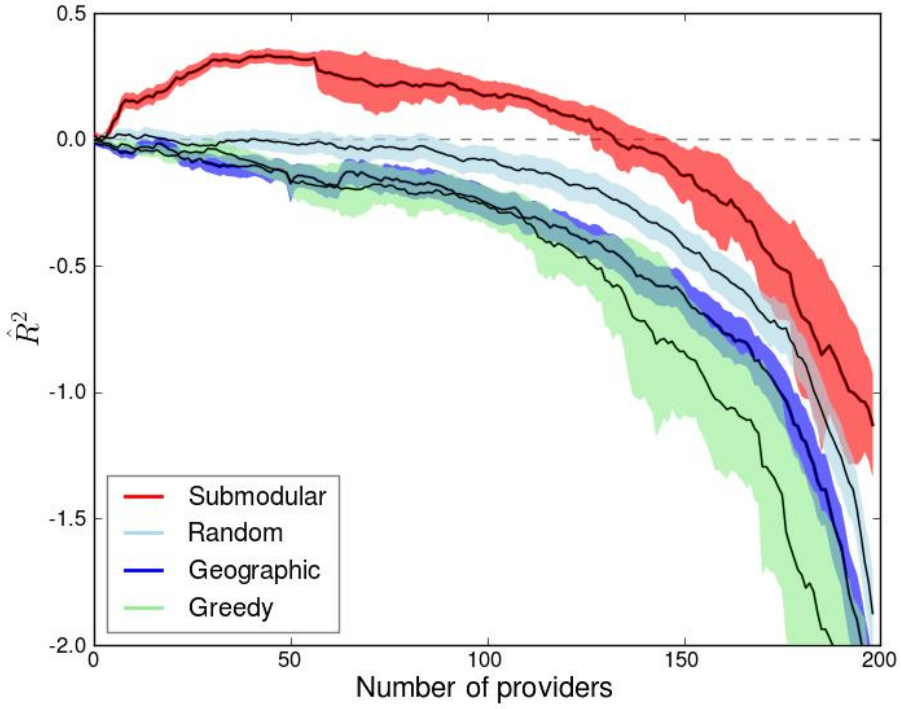
$$\hat{R}^2(\mathbf{G}^{2008}, S^{2007}, \tilde{\xi}) = \frac{\mathrm{Var}(\mathbf{G}^{2008}) - \mathrm{Var}(\mathbf{G}^{2008} - \sum_{i \in S} \alpha_i^{2007} \mathbf{P_i}^{2008}(\tilde{\xi}))}{\mathrm{Var}(\mathbf{G}^{2008})}.$$

Results: Out-of-sample Validation



$E_{\tilde{\xi}}[\hat{R}^2(\mathbf{G}^{2008}, S^{2007}, \tilde{\xi})]$ and 90% confidence intervals.

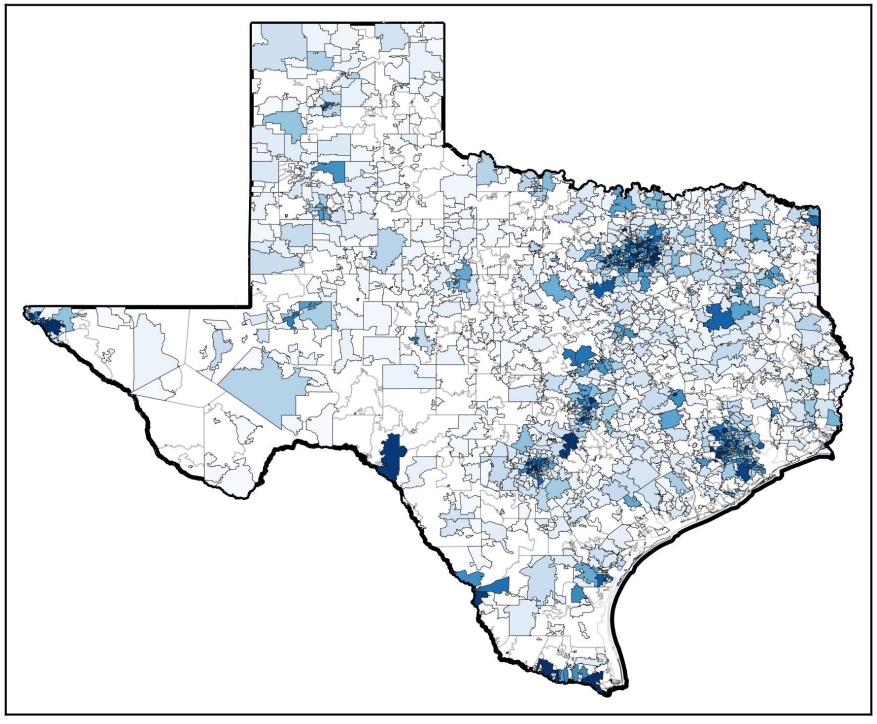
Results: Out-of-sample Validation



$E_{\tilde{\xi}}[\hat{R}^2(\mathbf{G}^{2008}, S^{2007}, \tilde{\xi})]$ and 90% confidence intervals.

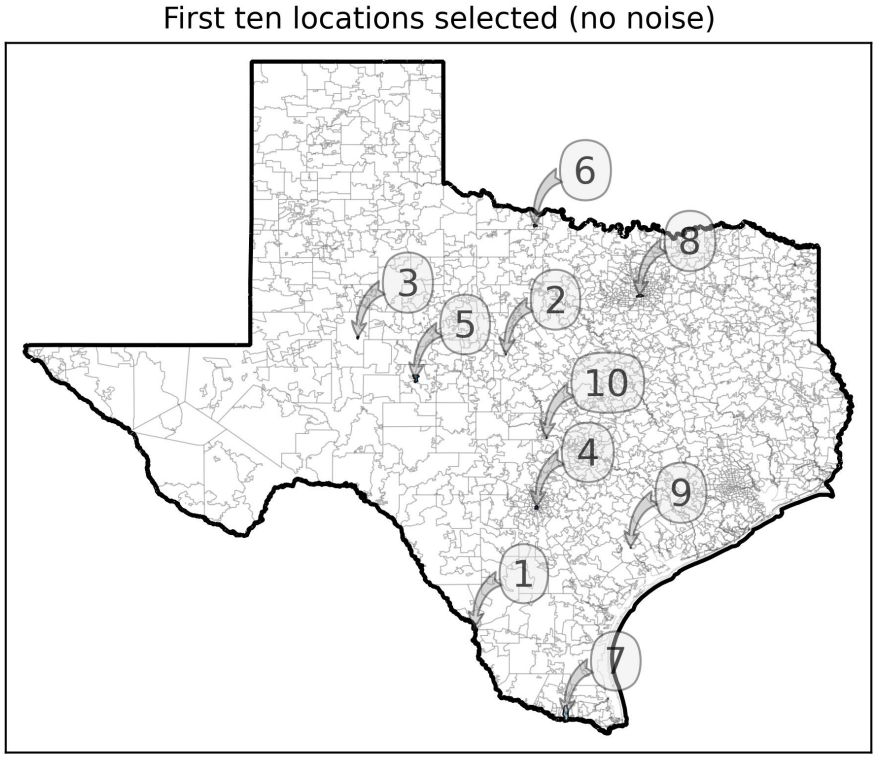
- Submodular is the only method with workable predictions.
- Adds noise after provider 60. (only 222 points in 2001-2007)

Results: Intuition



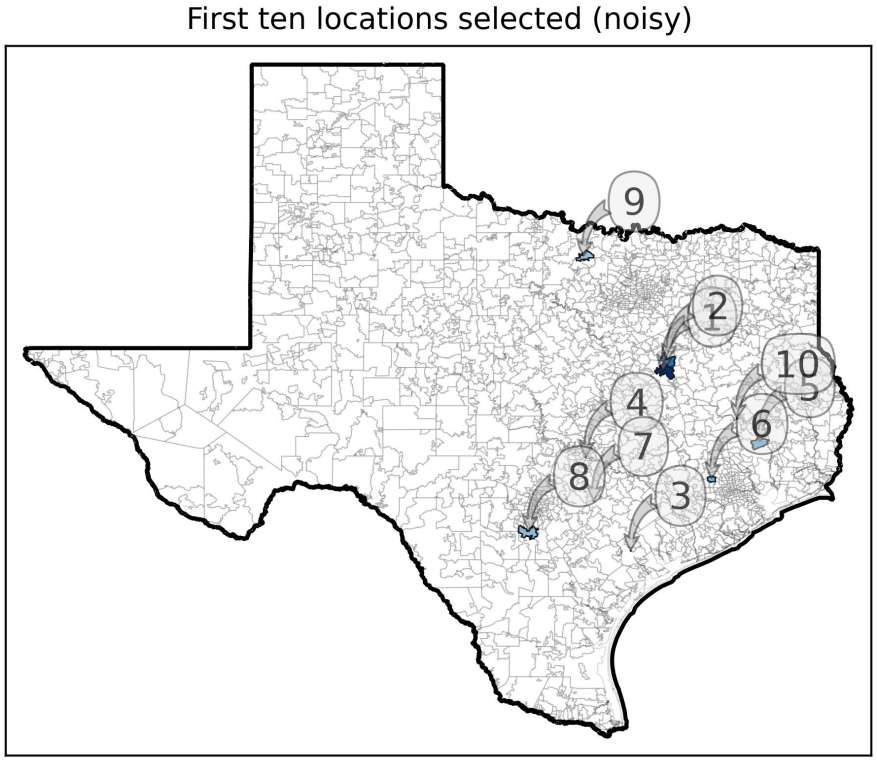
Texas population

Results: Intuition



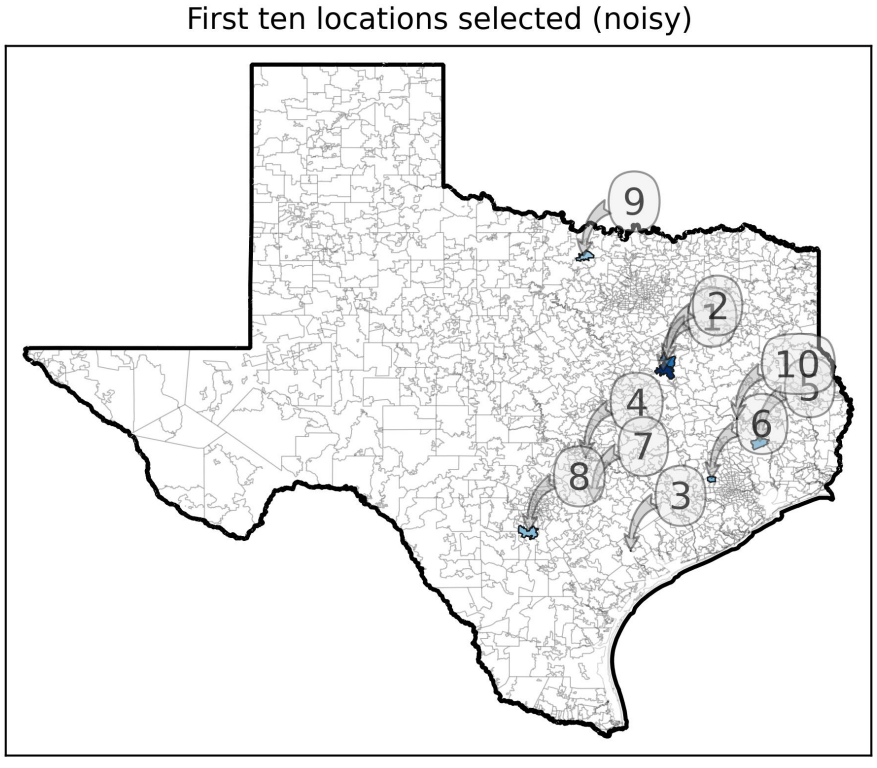
First 10 selections, no noise

Results: Intuition



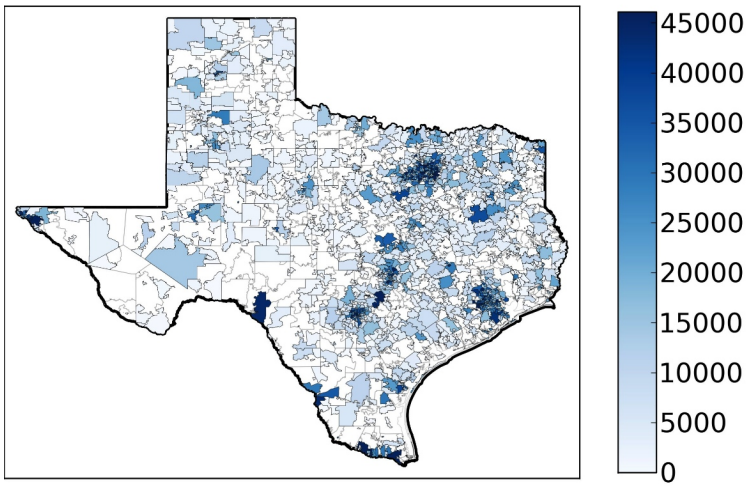
First 10 selections, with noise

Results: Intuition

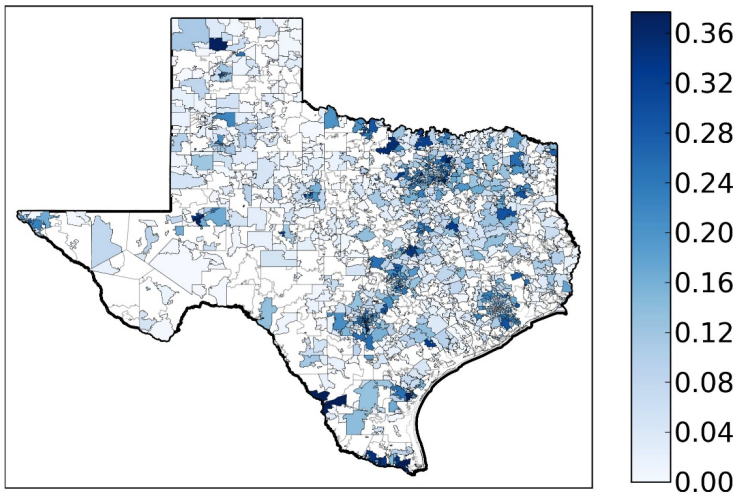


Lesson: Big cities behave the same when it comes to influenza.

Results: Intuition



Population



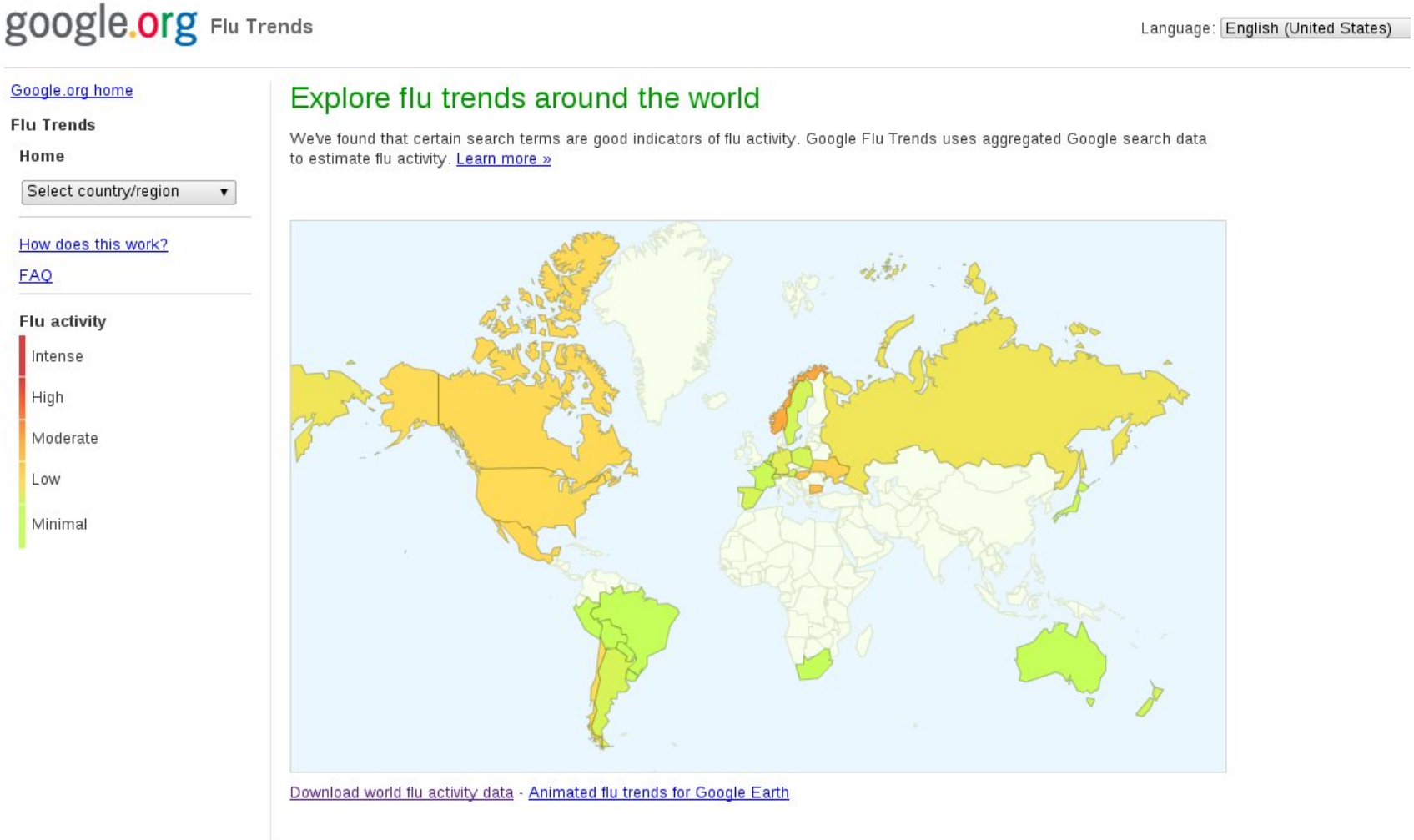
R^2 With First Principal Component

Centered, normalized hospitalization time series for PCA

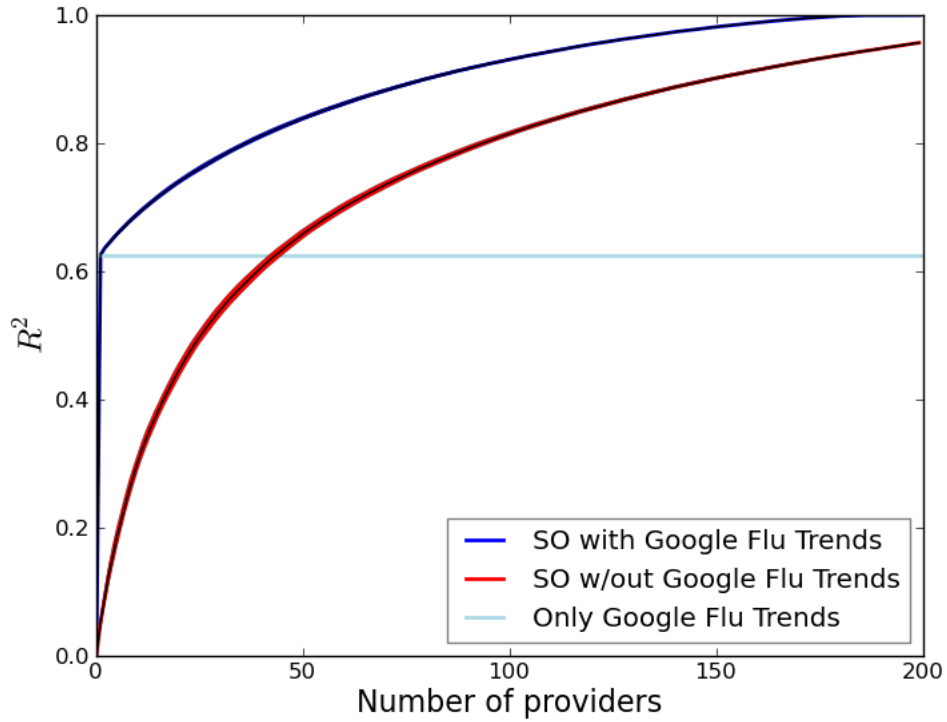
Results: New Technologies

A secondary question: Can Google Flu Trends substitute ILINet?

[Ginsberg et al. 08]

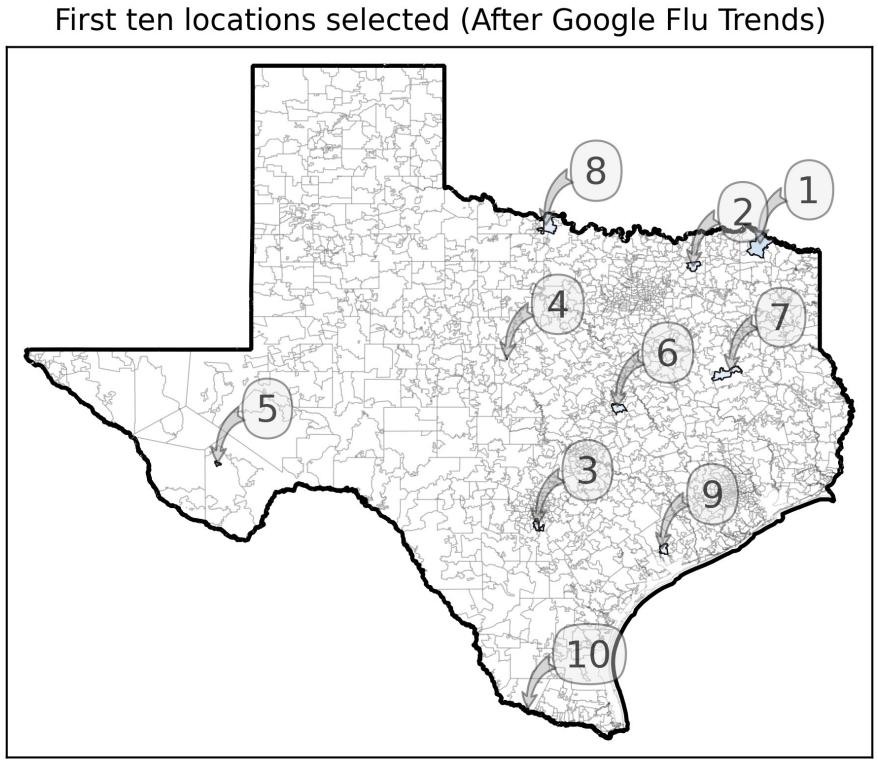


Results: New Technologies



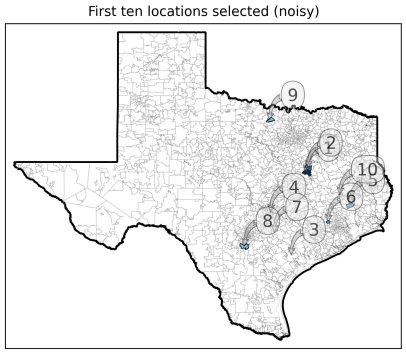
Google Flu Trends does very well, but ILINet also is useful.

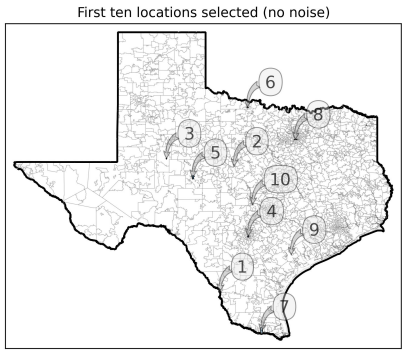
Results: New Technologies

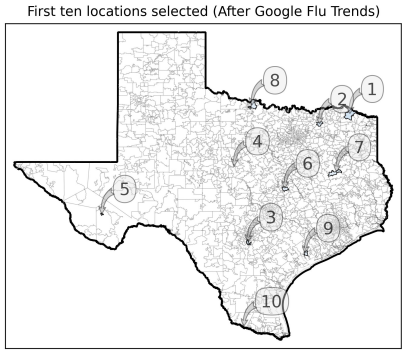


First 10 selections, with Google Flu Trends.

Results: New Technologies







Comparing first 10 providers in three networks.

Results: Pandemic Situations

Everything stops working in pandemic situations:

1. Google Flu Trends subject to panic
    - Out-of-sample for 2009 pandemic: e.g. 8k instead of 4k
    - Many more searches for flu-related terms

Results: Pandemic Situations

Everything stops working in pandemic situations:

1. Google Flu Trends subject to panic
   - Out-of-sample for 2009 pandemic: e.g. 8k instead of 4k
   - Many more searches for flu-related terms
2. Provider reports subject to panic
   - Out-of-sample for 2009 pandemic: e.g. 19k instead of 4k
   - Providers over-reporting?
   - Providers really start reporting?
   - More ILI patients go to the doctor?

Results: Pandemic Situations

Everything stops working in pandemic situations:

1. Google Flu Trends subject to panic
   - Out-of-sample for 2009 pandemic: e.g. 8k instead of 4k
   - Many more searches for flu-related terms
2. Provider reports subject to panic
   - Out-of-sample for 2009 pandemic: e.g. 19k instead of 4k
   - Providers over-reporting?
   - Providers really start reporting?
   - More ILI patients go to the doctor?
3. Noise model doesn't accurately capture panic

Results: Pandemic Situations

Focus on laboratory-confirmed cases of influenza:

- Also subject to panic, if doctors submit many more samples.
- Sequencing allows identification of new influenza strains.

Results: Pandemic Situations

Can re-do the optimizations with a new objective

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})] + w \cdot \text{Fraction of population covered by } S$$

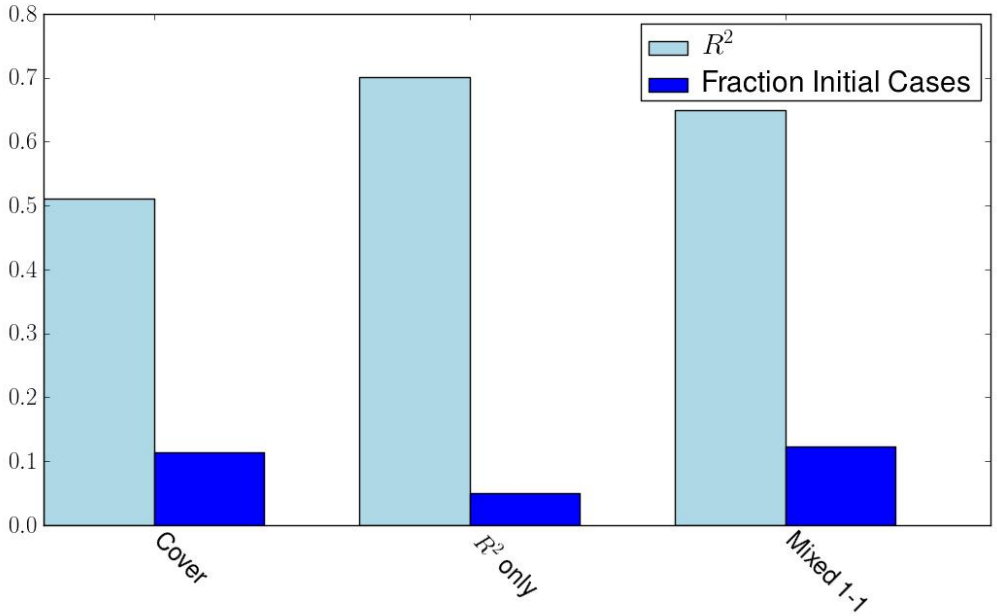
Results: Pandemic Situations

Can re-do the optimizations with a new objective

$$\max_{S \subseteq P} E_{\tilde{\xi}}[R^2(\mathbf{G}, S, \tilde{\xi})] + w \cdot \text{Fraction of population covered by } S$$

Submodularity still holds.

Results: Pandemic Situations



Can create networks that both:

1. Have good R^2 in non-pandemic situations
2. Capture a good fraction of cases for lab-confirmations

Summary and Conclusions

Designing the Texas Influenza-like-illness Network

- Results used by TX DSHS
- Applicable to other geographic regions
- Applicable to other objectives
- Intuitive conclusion

Summary and Conclusions

Designing the Texas Influenza-like-illness Network

- Results used by TX DSHS
- Applicable to other geographic regions
- Applicable to other objectives
- Intuitive conclusion

Different diseases/objectives could produce different networks.

Summary and Conclusions

Method requires historical data:

1. Goal time series
2. Provider noise profiles
3. Estimates of ILI cases per geographic region

- Often available for developed countries
- Perhaps can use alternate sources for developing countries

Summary and Conclusions

Interesting directions:

- Pandemic situations
- Data-poor situations
- Surveillance for other diseases

Thank you.

Proof of Greedy Submodular Approximation

- S^* - optimal solution with k items
- S^G - greedy solution with j items
- ρ_i - the bump in greedy solution on ith iteration

$$f(S^*) \leq f(S^G) + \sum_{x \in S^* - S^G} [f(S^G + x) - f(S^G)]$$

$$f(S^*) \leq \sum_{i=1}^{j} \rho_i + k \cdot \rho_{j+1}$$

Proof of Greedy Submodular Approximation

$$f(S^*) \leq f(S^G) + \sum_{x \in S^* - S^G} [f(S^G + x) - f(S^G)]$$

$$f(S^*) \leq \sum_{i=1}^{j} \rho_i + k \cdot \rho_{j+1}$$

$$\rho_{j+1} \geq \frac{1}{k} \cdot f(S^*) - \frac{1}{k} \sum_{i=1}^{j} \rho_i$$

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{1}{k} \cdot f(S^*) - \frac{k-1}{k} \sum_{i=1}^{j} \rho_i$$

Proof of Greedy Submodular Approximation

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{1}{k} \cdot f(S^*) - \frac{k-1}{k} \sum_{i=1}^{j} \rho_i$$

Induction:

$$\sum_{i=1}^{j} \rho_i \geq \frac{k^j - (k-1)^j}{k^j} \cdot f(S^*)$$

Base:

$$\rho_1 \geq \frac{1}{k} \cdot f(S^*)$$

Proof of Greedy Submodular Approximation

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{1}{k} \cdot f(S^*) - \frac{k-1}{k} \sum_{i=1}^{j} \rho_i$$

Induction:

$$\sum_{i=1}^{j} \rho_i \geq \frac{k^j - (k-1)^j}{k^j} \cdot f(S^*)$$

Induction Step (prove for $j+1$ given statement for j):

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{1}{k} \cdot f(S^*) - \frac{k-1}{k} \cdot \frac{k^j - (k-1)^j}{k^j} \cdot f(S^*)$$

Proof of Greedy Submodular Approximation

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{1}{k} \cdot f(S^*) - \frac{k-1}{k} \sum_{i=1}^{j} \rho_i$$

Induction:

$$\sum_{i=1}^{j} \rho_i \geq \frac{k^j - (k-1)^j}{k^j} \cdot f(S^*)$$

Induction Step (prove for $j+1$ given statement for j):

$$\sum_{i=1}^{j+1} \rho_i \geq \frac{k^{j+1} - (k-1)^{j+1}}{k^{j+1}} \cdot f(S^*)$$

Proof of Greedy Submodular Approximation

$$\sum_{i=1}^{j} \rho_i \geq \frac{k^j - (k-1)^j}{k^j} \cdot f(S^*)$$

Substitute $j = k$ to get:

$$f(S^G) = \sum_{i=1}^{k} \rho_i \geq 1 - (1 - \frac{1}{k})^k \cdot f(S^*) \geq (1 - \frac{1}{e}) f(S^*)$$